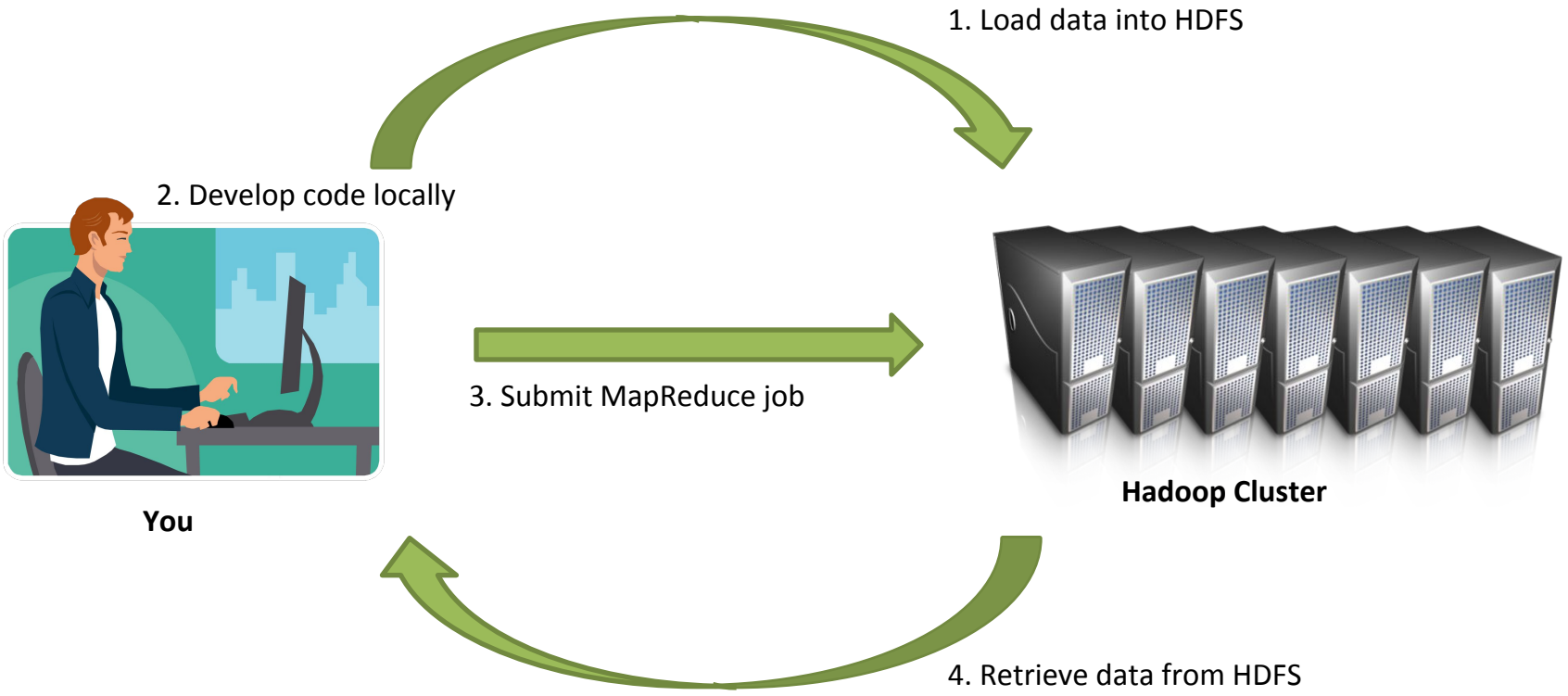
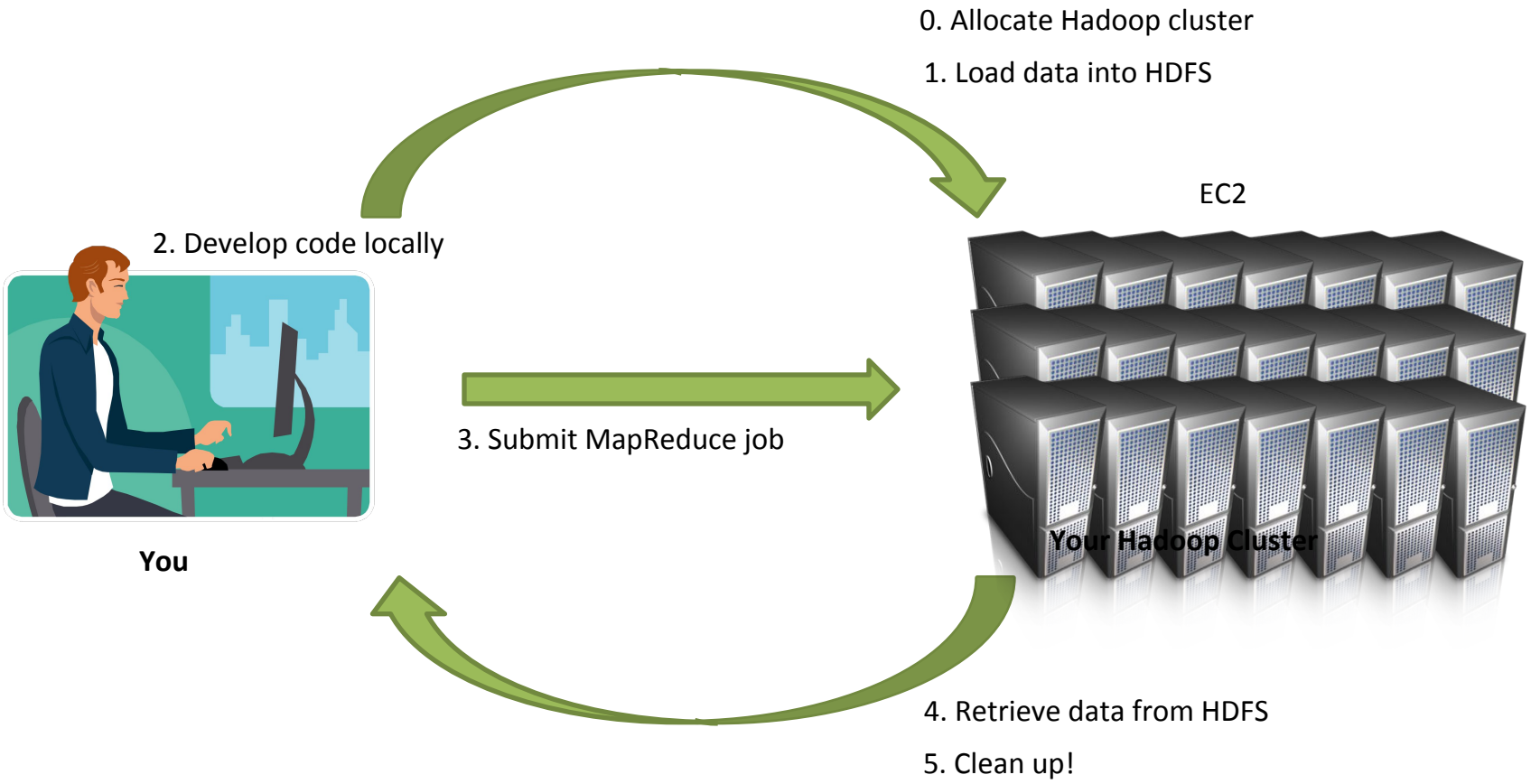
MapReduce

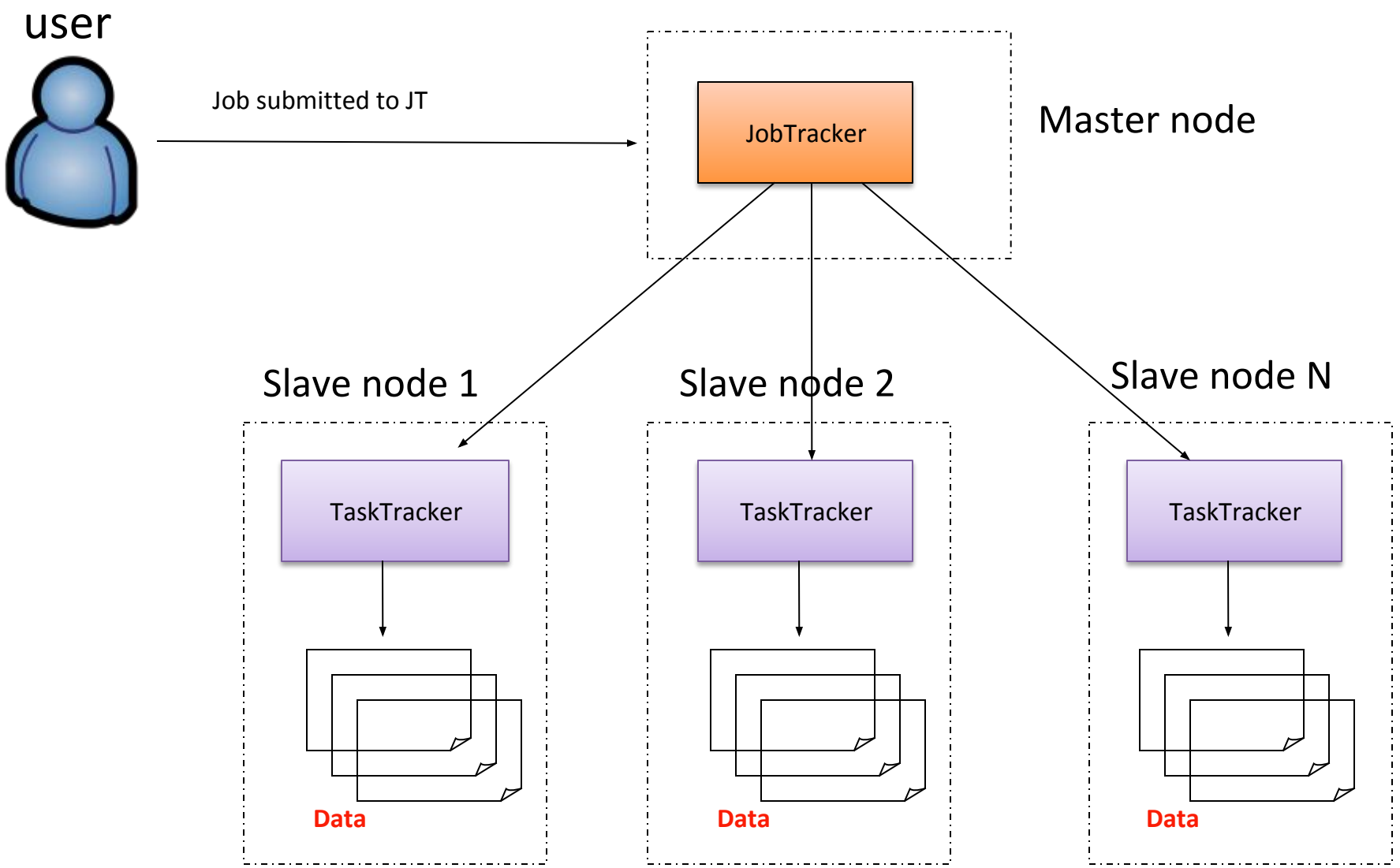
Hadoop Workflow

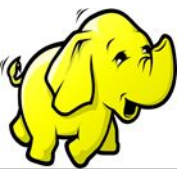


On Amazon: With EC2

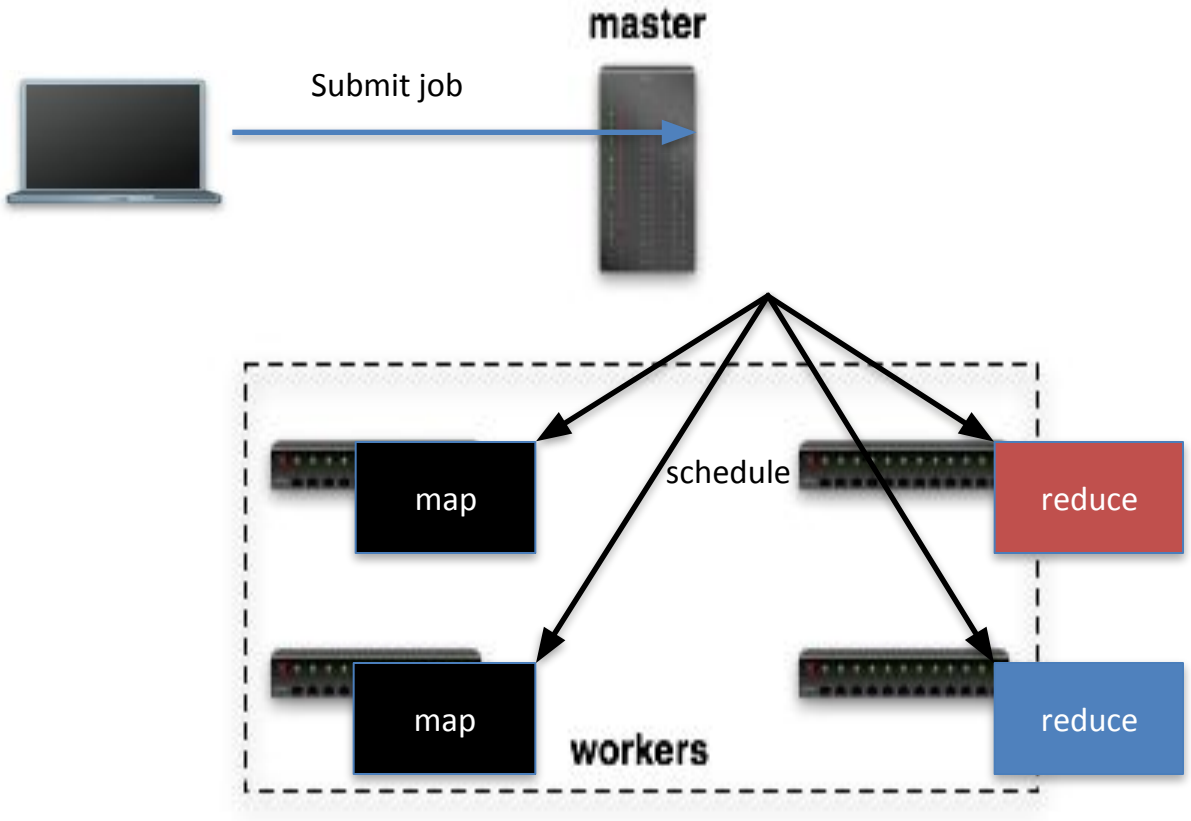


MapReduce Architecture overview

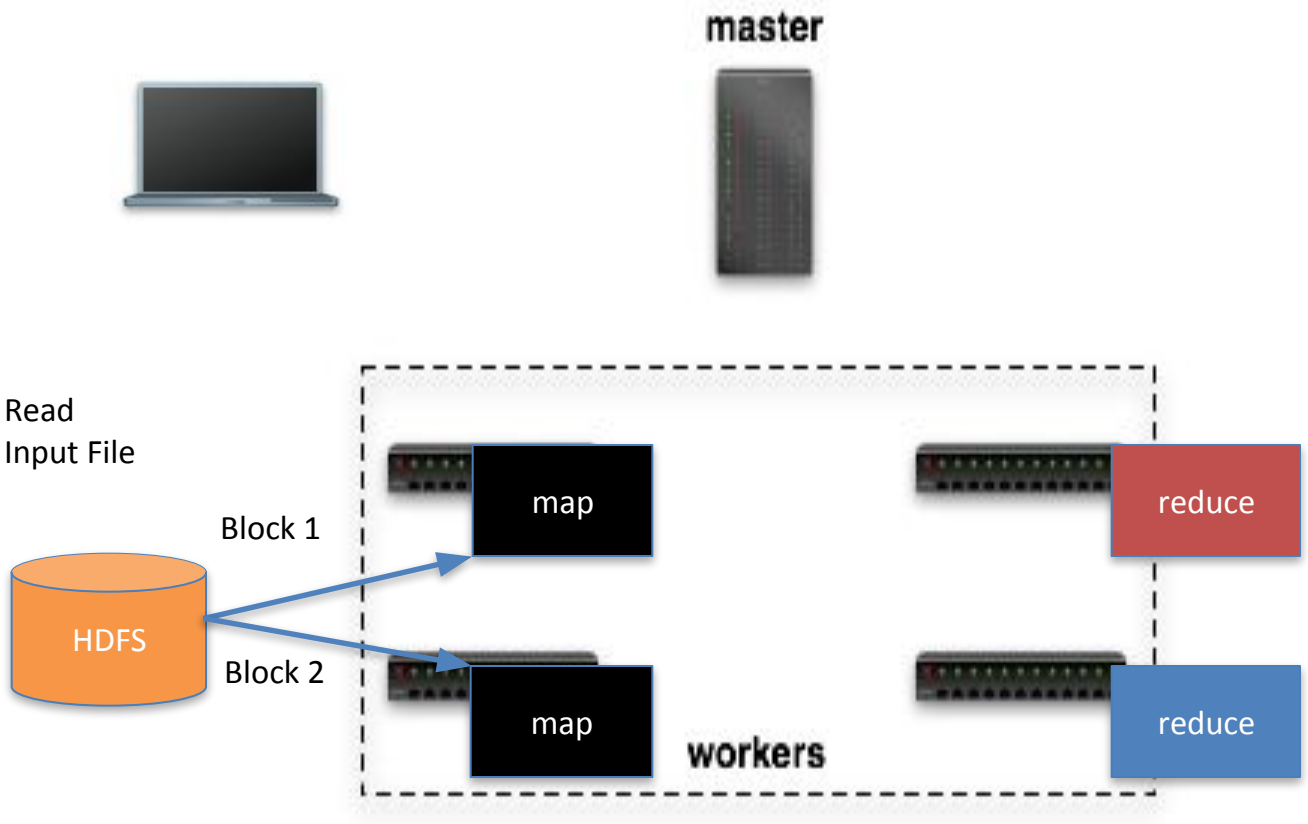


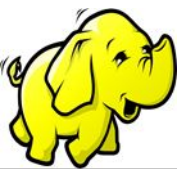


Dataflow in Hadoop

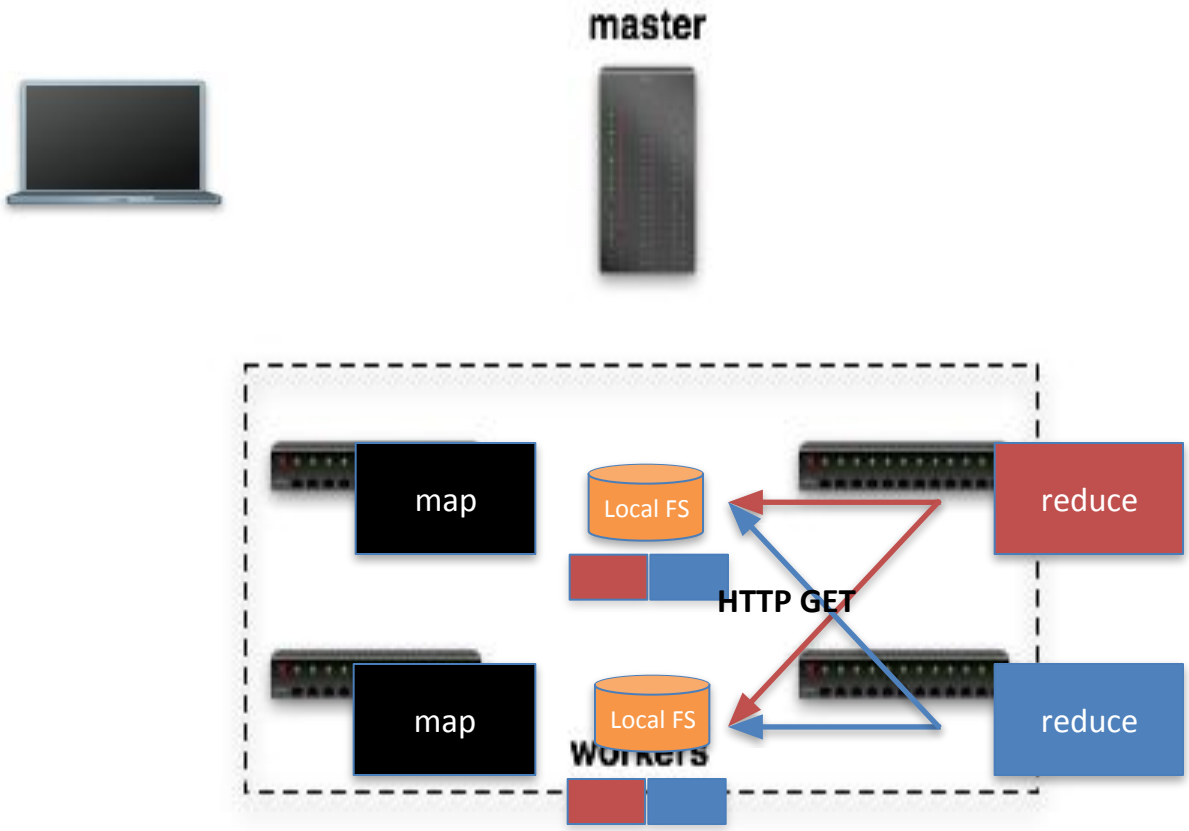


Dataflow in Hadoop

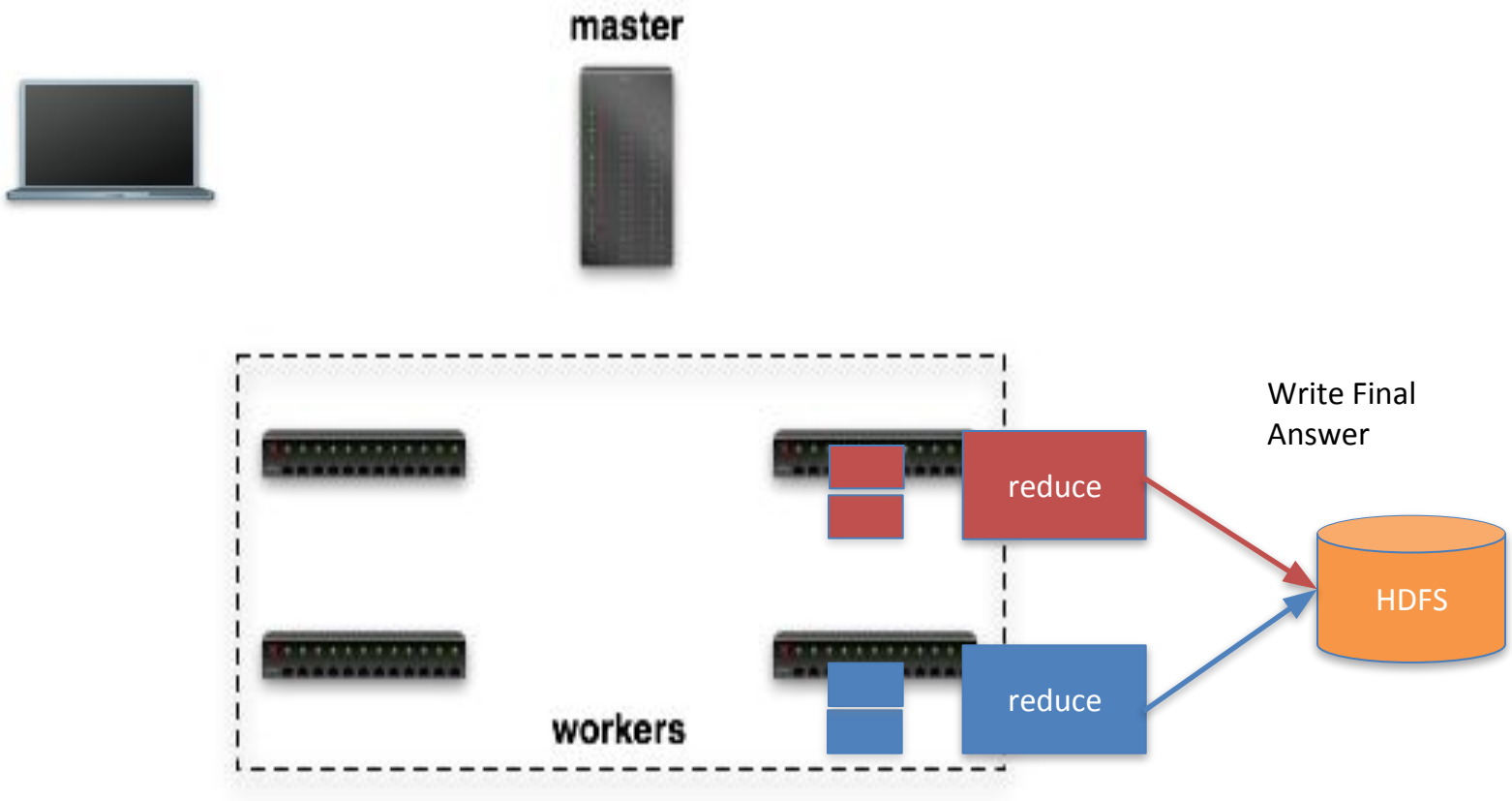




Dataflow in Hadoop

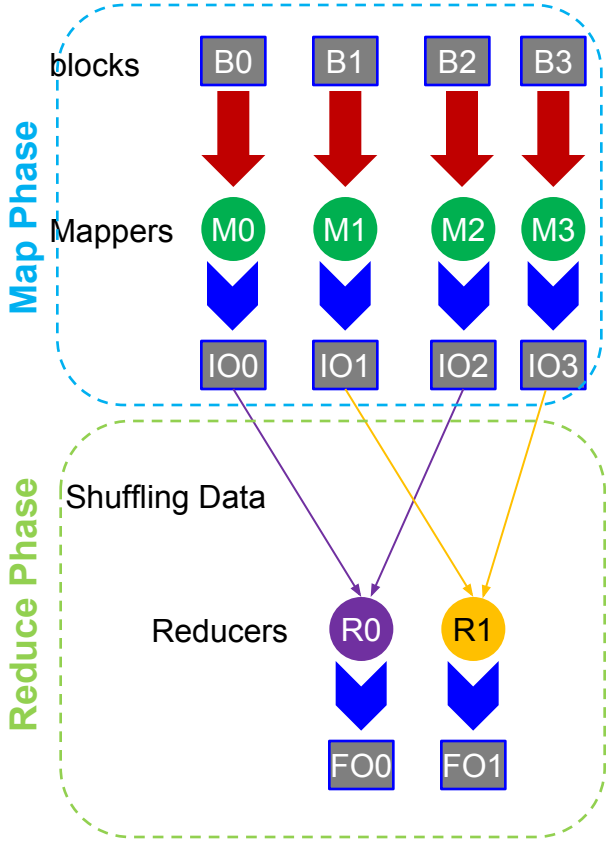


Dataflow in Hadoop



MapReduce: A Bird's-Eye View

- In MapReduce, blocks are processed in isolation by tasks called *Mappers*
- The outputs from the mappers are denoted as intermediate outputs (IOs) and are brought into a second set of tasks called *Reducers*
- The process of bringing together IOs into a set of Reducers is known as *shuffling process*
- The Reducers produce the final outputs (FOs)

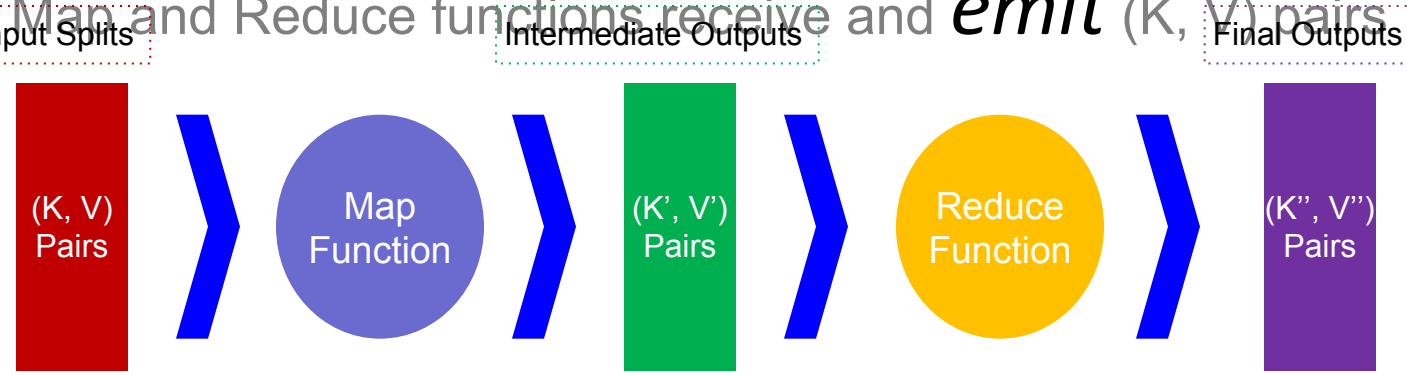


Keys and Values

- The programmer in MapReduce has to specify two functions, the *Map function* and the *Reduce function* that implement the Mapper and the Reducer in a MapReduce program

- In MapReduce data elements are always structured as key-value (i.e., (K, V)) pairs

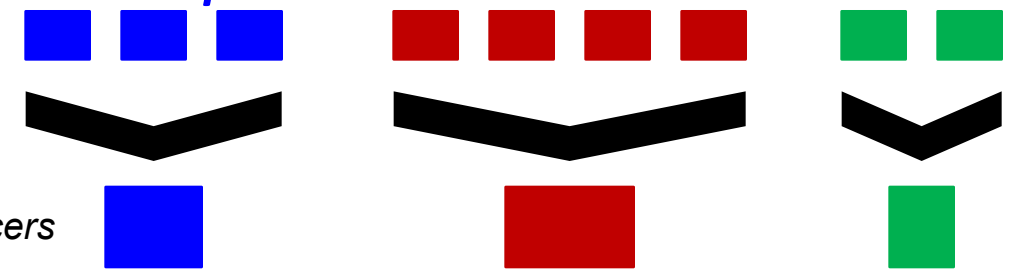
- The Map and Reduce functions receive and *emit* (K, V) pairs



Partitions

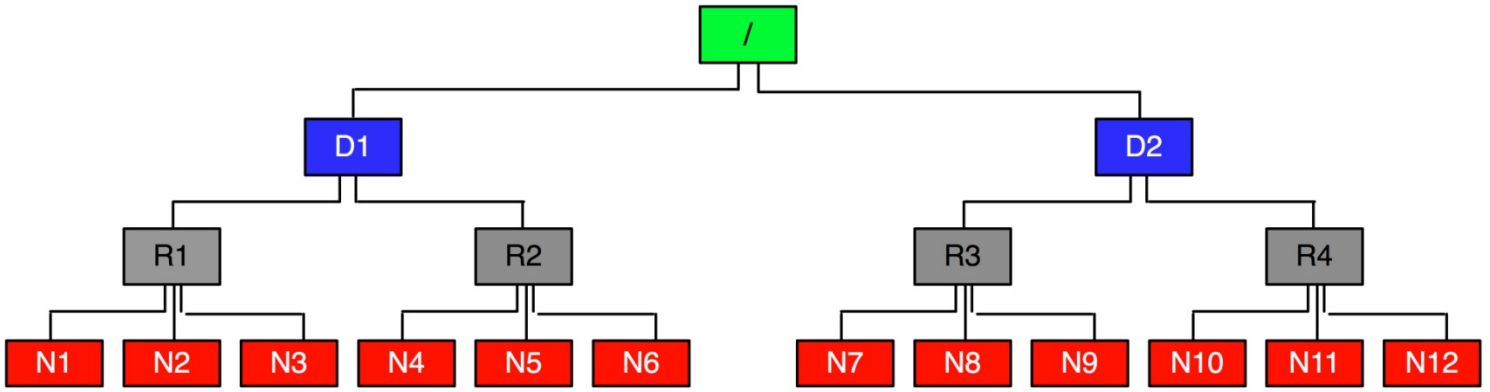
- In MapReduce, intermediate output values are not usually reduced together
- *All values with the same key are presented to a single Reducer together*
- More specifically, a different subset of intermediate key space is assigned to each Reducer
- These subsets are known as *partitions*

Different colors represent different keys (potentially) from different Mappers



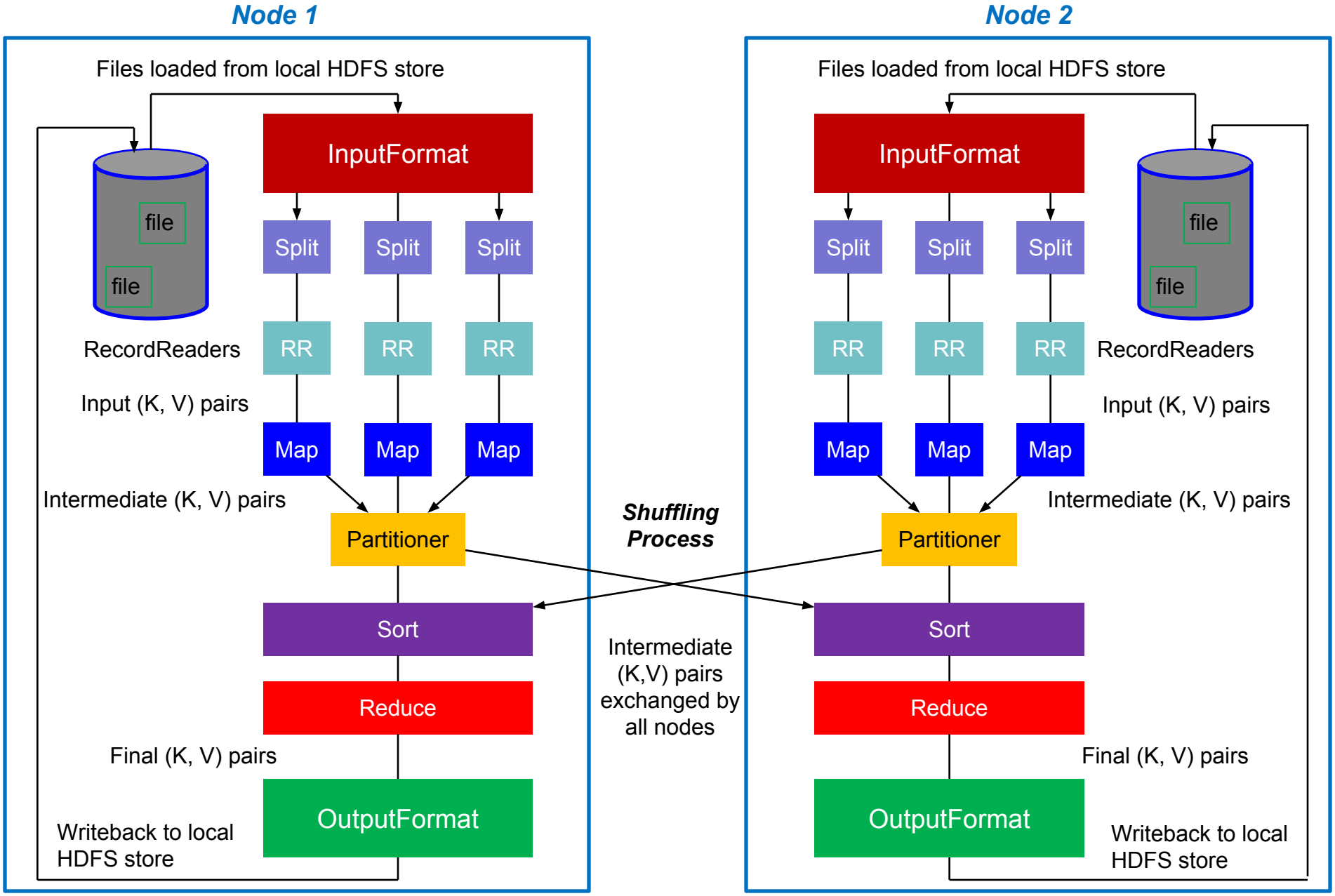
Partitions are the input to Reducers

Network Topology In MapReduce



- MapReduce assumes a tree style network topology
- Nodes are spread over different racks embraced in one or many data centers
- A salient point is that the bandwidth between two nodes is dependent on their relative locations in the network topology

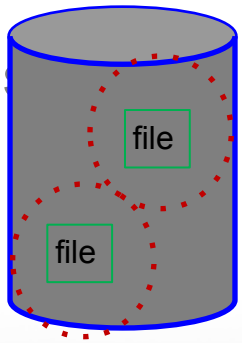
Hadoop MapReduce: A Closer Look



Input Files

- *Input files* are where the data for a MapReduce task is initially stored

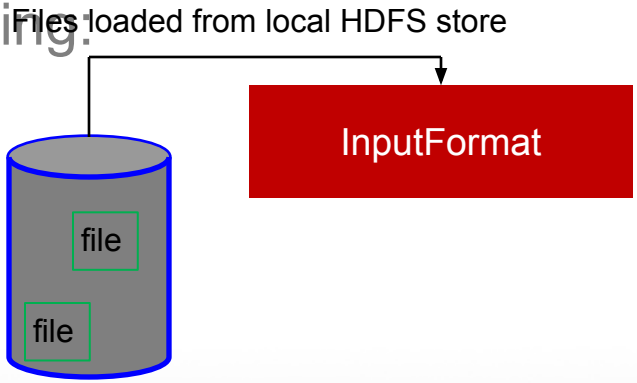
- The input files typically reside in a distributed file system (e.g. HDFS)



- The format of input files is arbitrary
  - Line-based log files
  - Binary files
  - Multi-line input records, etc.

InputFormat

- How the input files are split up and read is defined by the *InputFormat*
- InputFormat is a class that does the following:
  - Selects the files that should be used for input
  - Defines the *InputSplits* that break a file
  - Provides a factory for *RecordReader* objects that read the file



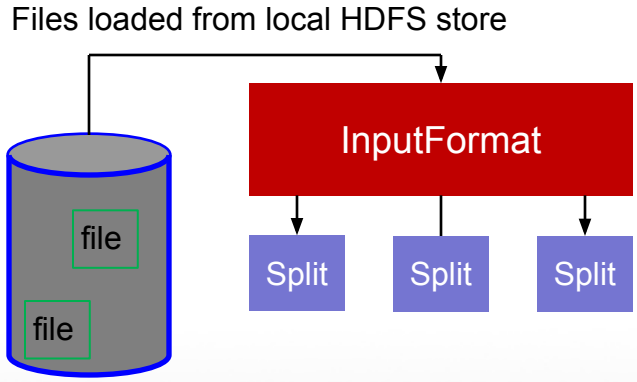
InputFormat Types

- Several InputFormats are provided with Hadoop:

InputFormat	Description	Key	Value
TextInputFormat	Default format; reads lines of text files	The byte offset of the line	The line contents
KeyValueInputFormat	Parses lines into (K, V) pairs	Everything up the first tab character	
SequenceFileInputFormat	A Hadoop-specific high-performance binary format	user-defined	user-defined

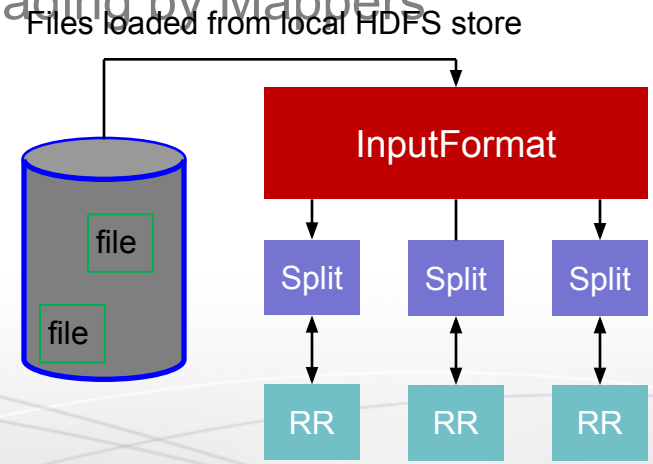
Input Splits

- An *input split* describes a unit of work that comprises a single map task in a MapReduce program
- By default, the InputFormat breaks a file up into 64MB splits
- By dividing the file into splits, we allow several map tasks to operate on a single file in parallel
- If the file is very large, this can improve performance significantly through parallelism



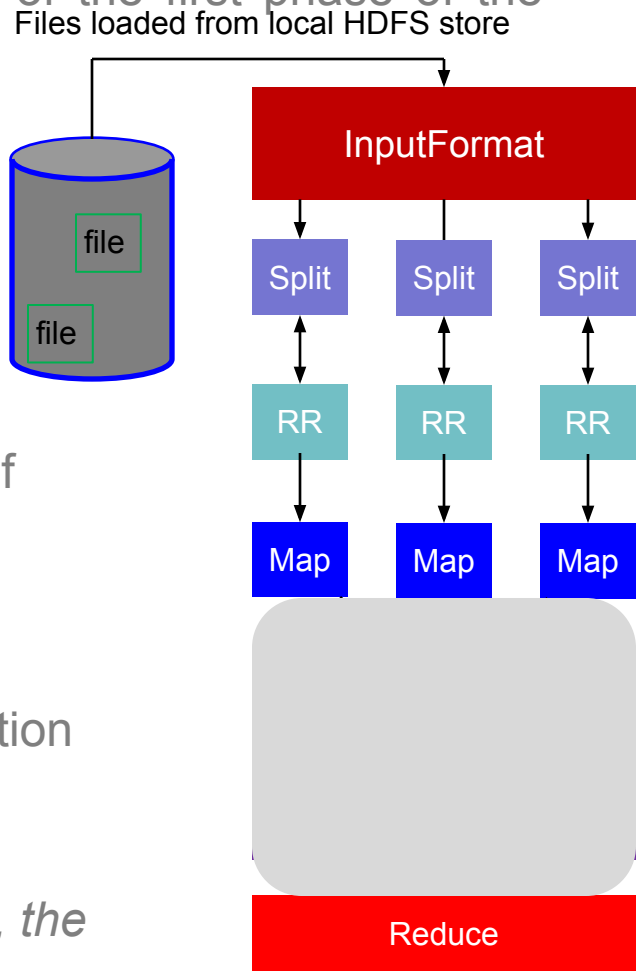
RecordReader

- The input split defines a slice of work but does not describe how to access it
- The *RecordReader* class actually loads data from its source and converts it into (K, V) pairs suitable for reading by Mappers
- The RecordReader is invoked repeatedly on the input until the entire split is consumed
- Each invocation of the RecordReader leads to another call of the map function defined by the programmer



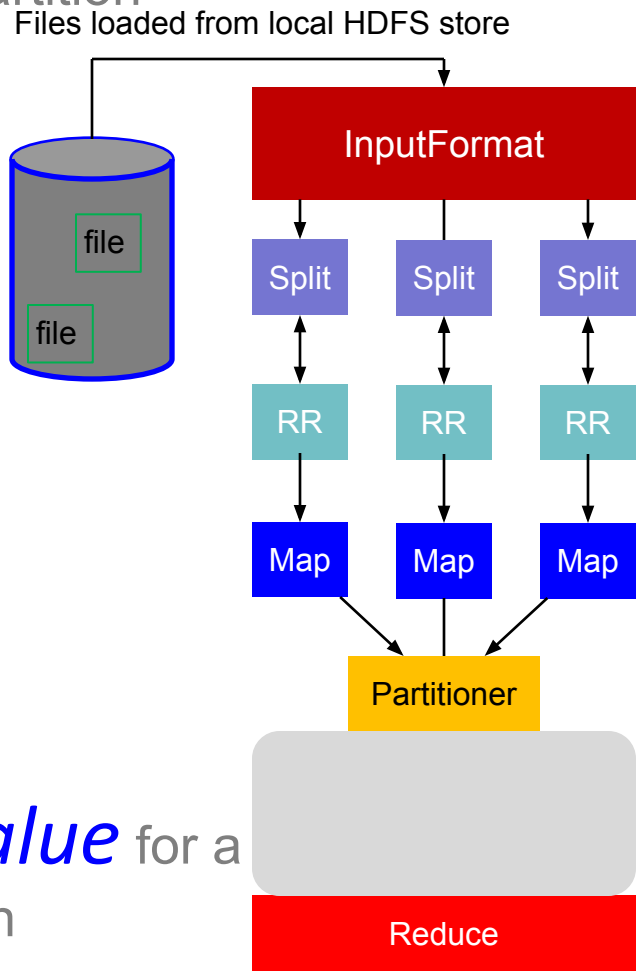
Mapper and Reducer

- The *Mapper* performs the user-defined work of the first phase of the MapReduce program
- A new instance of Mapper is created for each split
- The *Reducer* performs the user-defined work of the second phase of the MapReduce program
- A new instance of Reducer is created for each partition
- *For each key in the partition assigned to a Reducer, the Reducer is called once*



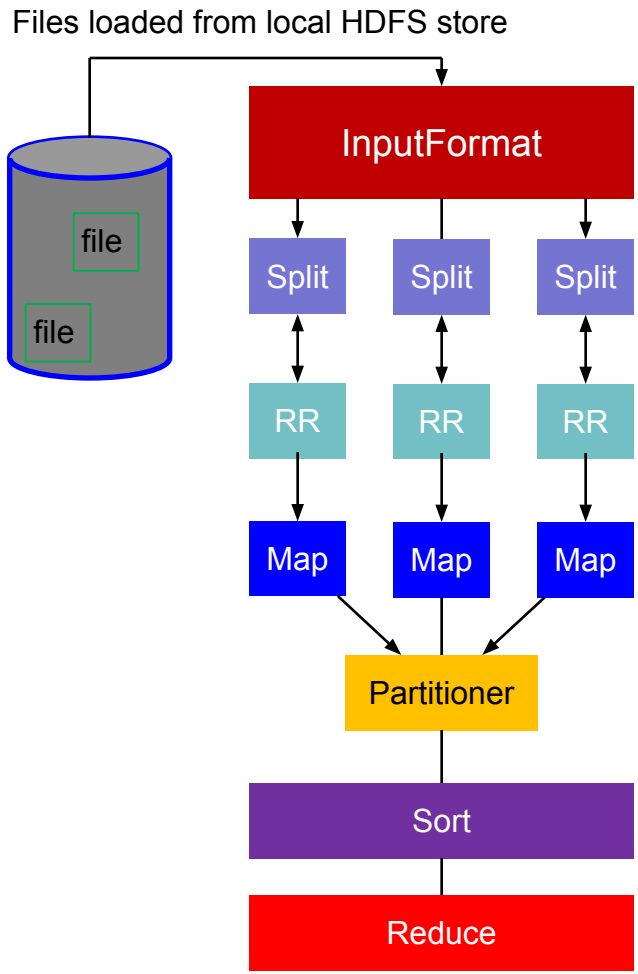
Partitioner

- Each mapper may emit (K, V) pairs to *any* partition
- Therefore, the map nodes must all agree on where to send different pieces of intermediate data
- The *partitioner* class determines which partition a given (K,V) pair will go to
- The default partitioner computes *a hash value* for a given key and assigns it to a partition based on this result



Sort

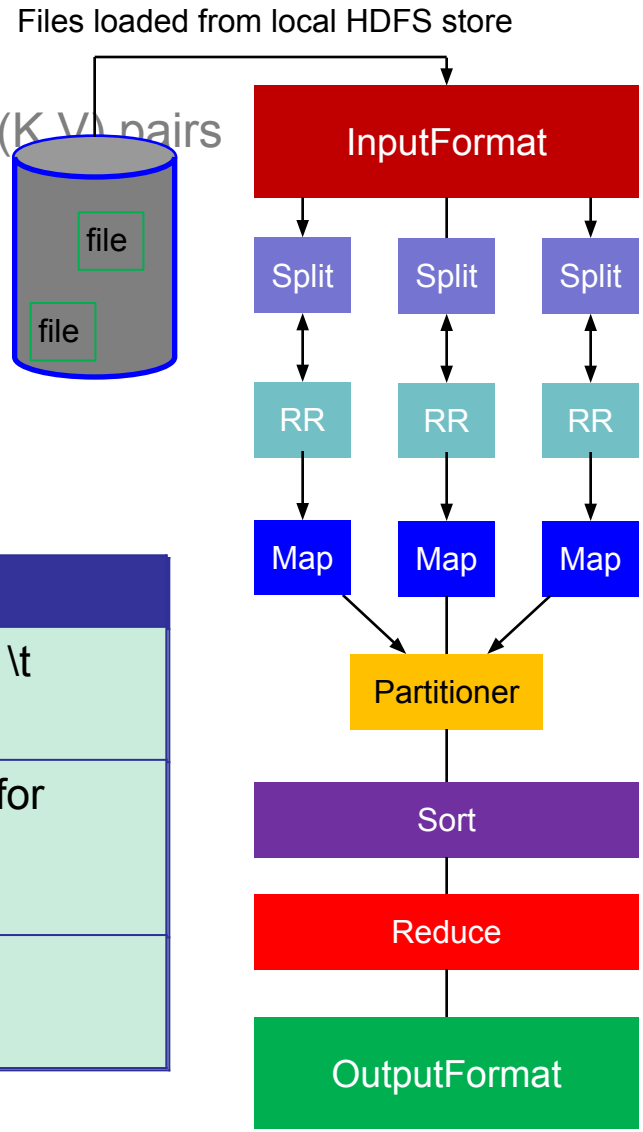
- Each Reducer is responsible for reducing the values associated with (several) intermediate keys
- The set of intermediate keys on a single node is *automatically sorted* by MapReduce before they are presented to the Reducer



OutputFormat

- The *OutputFormat* class defines the way (K,V) pairs produced by Reducers are written to output files

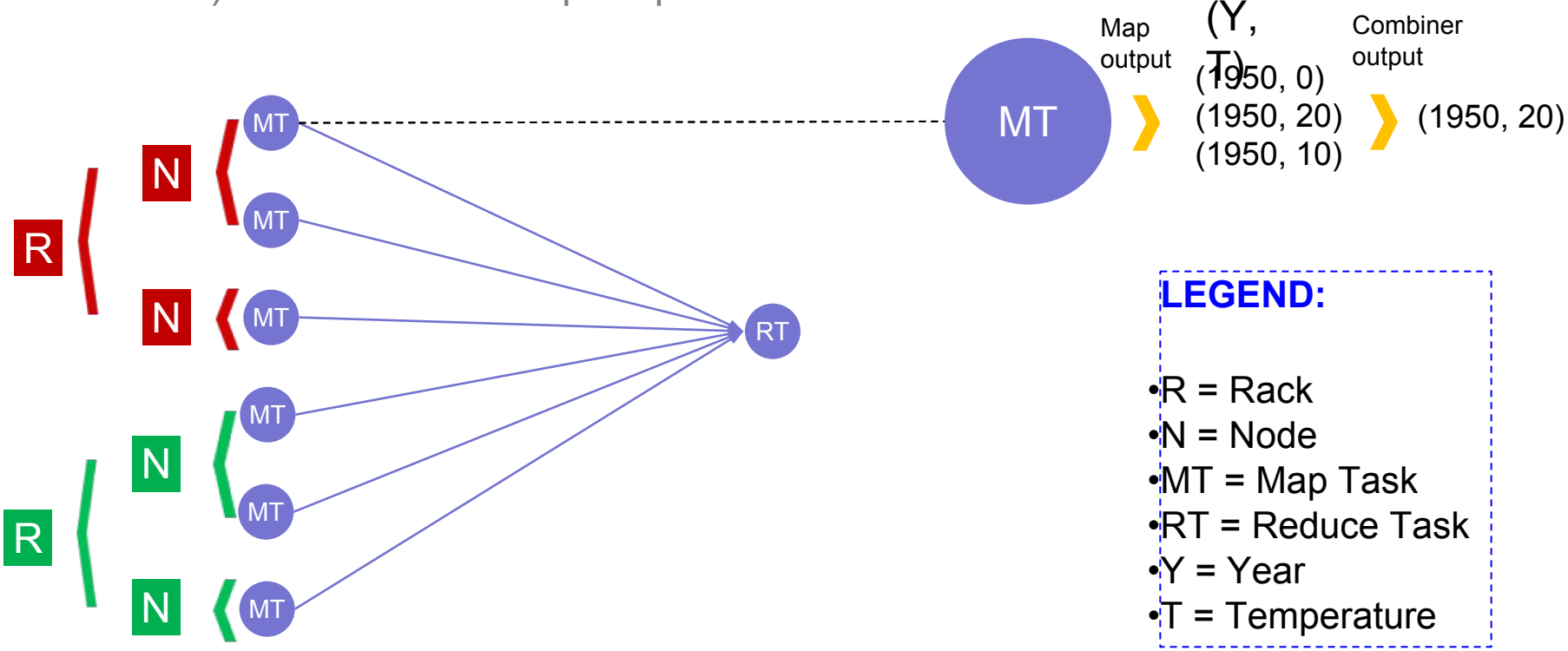
- The instances of OutputFormat provided by Hadoop write to files on the local disk or in HDFS



OutputFormat	Description
TextOutputFormat	Default; writes lines in "key \t value" format
SequenceFileOutputFormat	Writes binary files suitable for reading into subsequent MapReduce jobs
NullOutputFormat	Generates no output files

Combiner Functions

- MapReduce applications are limited by the bandwidth available on the cluster
- It pays to minimize the data shuffled between map and reduce tasks
- Hadoop allows the user to specify a combiner function (just like the reduce function) to be run on a map output

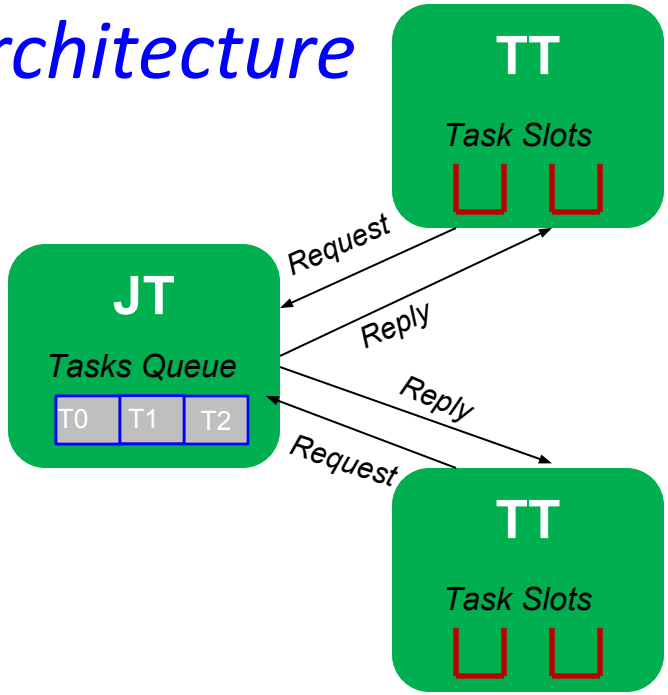


Job Scheduling in MapReduce

- In MapReduce, an application is represented as a *job*
- A job encompasses multiple map and reduce tasks
- MapReduce in Hadoop comes with a choice of schedulers:
  - The default is the *FIFO scheduler* which schedules jobs in order of submission
  - There is also a multi-user scheduler called the *Fair scheduler* which aims to give every user a fair share of the cluster
    capacity over time

Task Scheduling in MapReduce

- MapReduce adopts a *master-slave architecture*
- The master node in MapReduce is referred to as *Job Tracker* (JT)
- Each slave node in MapReduce is referred to as *Task Tracker* (TT)



- MapReduce adopts a *pull scheduling* strategy rather than a *push one*

Map and Reduce Task Scheduling

- Every TT sends a *heartbeat message* periodically to JT encompassing a request for a map or a reduce task to run

I. Map Task Scheduling:

- JT satisfies requests for map tasks via attempting to schedule mappers in the *vicinity* of their input splits (i.e., it considers locality)

II. Reduce Task Scheduling:

- However, JT simply assigns the next yet-to-run reduce task to a requesting TT regardless of TT's network location and its implied effect on the reducer's shuffle time (i.e., it does not consider locality)

Fault Tolerance in MapReduce

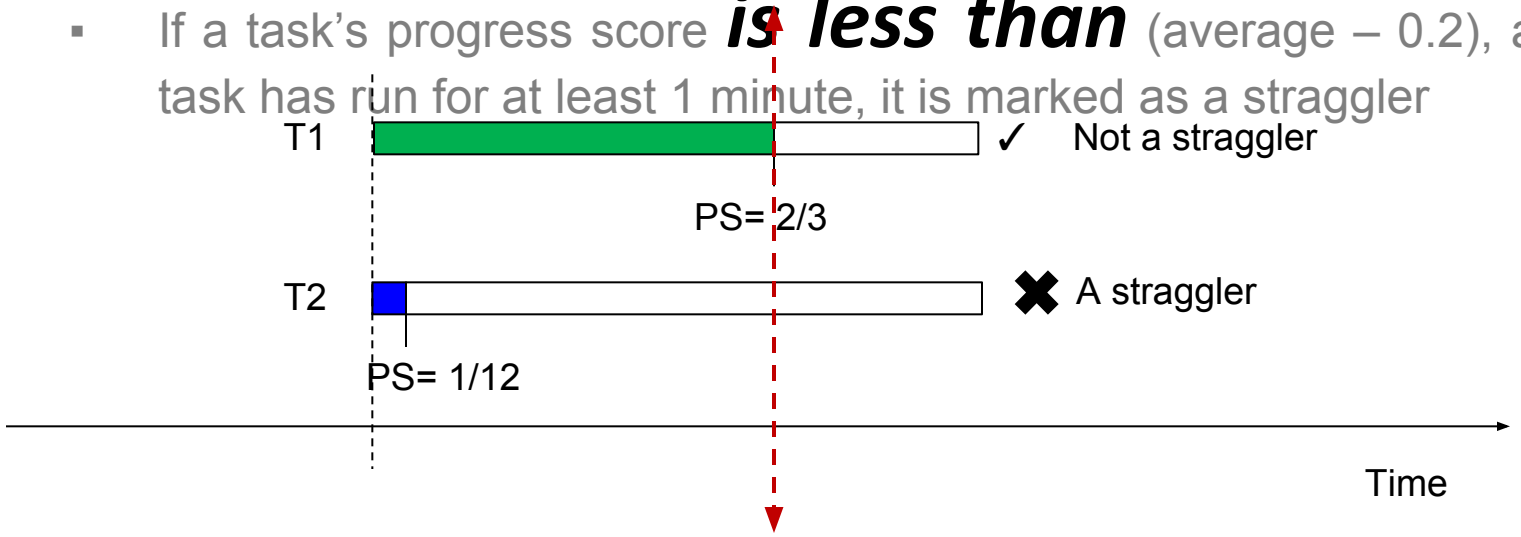
- MapReduce can guide jobs toward a successful completion even when jobs are run on a large cluster where probability of failures increases
- The primary way that MapReduce achieves fault tolerance is through *restarting tasks*
- If a TT fails to communicate with JT for a period of time, JT will assume that TT in question has crashed
  - If the job is still in the map phase, JT asks another TT to re-execute ***all Mappers that previously ran at the failed TT***
  - If the job is in the reduce phase, JT asks another TT to re-execute ***all***

Speculative Execution

- A MapReduce job is dominated by the slowest task
- MapReduce attempts to locate slow tasks (*stragglers*) and run redundant (*speculative*) tasks that will optimistically commit before the corresponding stragglers
- This process is known as *speculative execution*
- Only one copy of a straggler is allowed to be speculated
- Whichever copy (among the two copies) of a task commits first, it

Locating Stragglers

- How does Hadoop locate stragglers?
  - Hadoop monitors each task progress using a *progress score* between 0 and 1
  - If a task's progress score ***is less than*** (average – 0.2), and the task has run for at least 1 minute, it is marked as a straggler



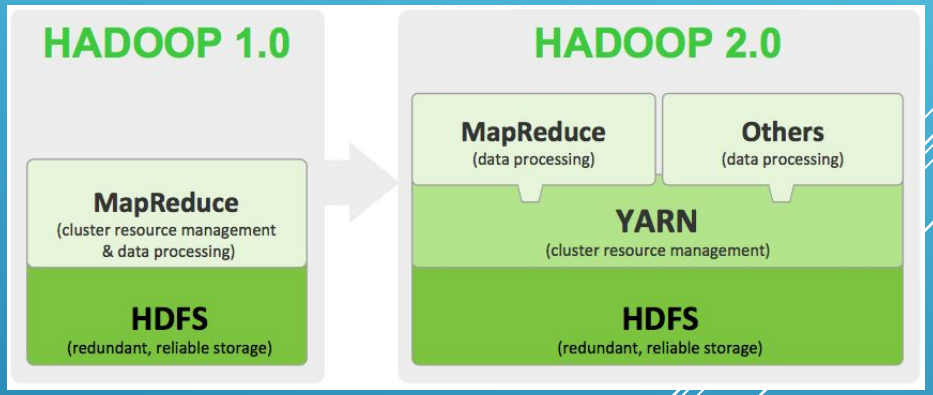
YARN

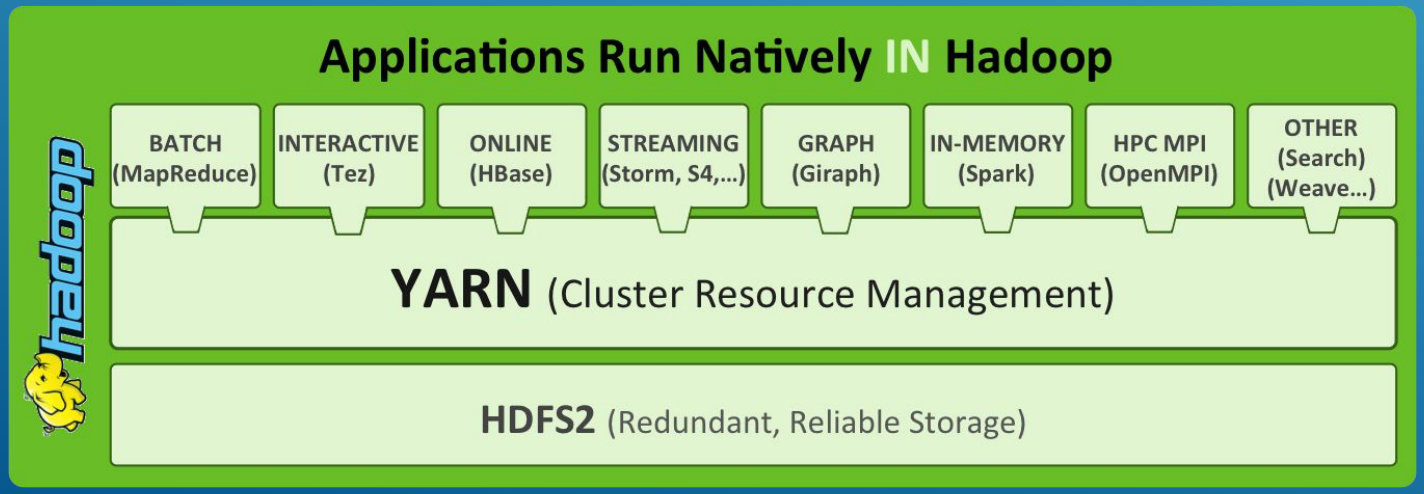
YARN

Sometimes called MapReduce 2.0, YARN decouples scheduling capabilities from the data processing component

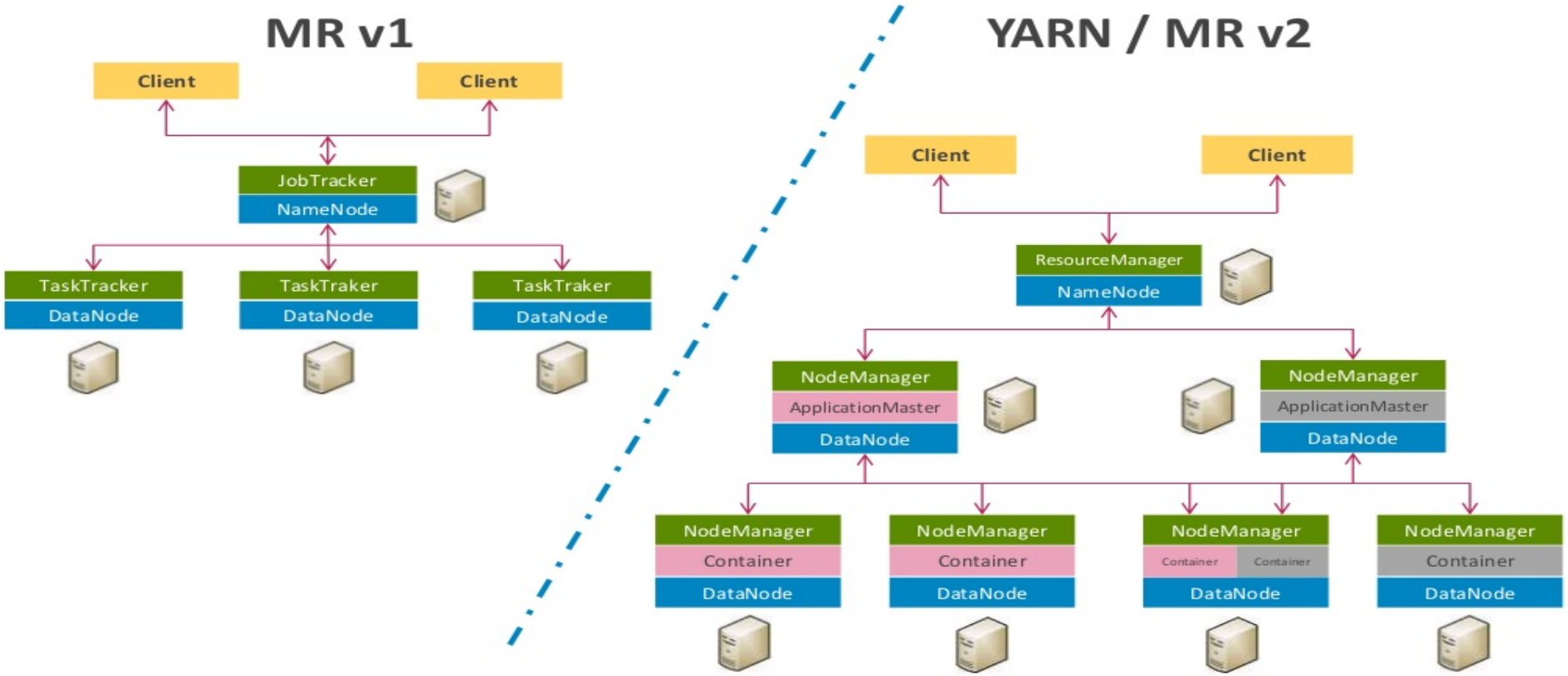
Hadoop clusters can now run interactive querying and streaming data applications simultaneously.

Separating HDFS from MapReduce with YARN makes the Hadoop environment more suitable for operational applications that can't wait for batch jobs to finish.



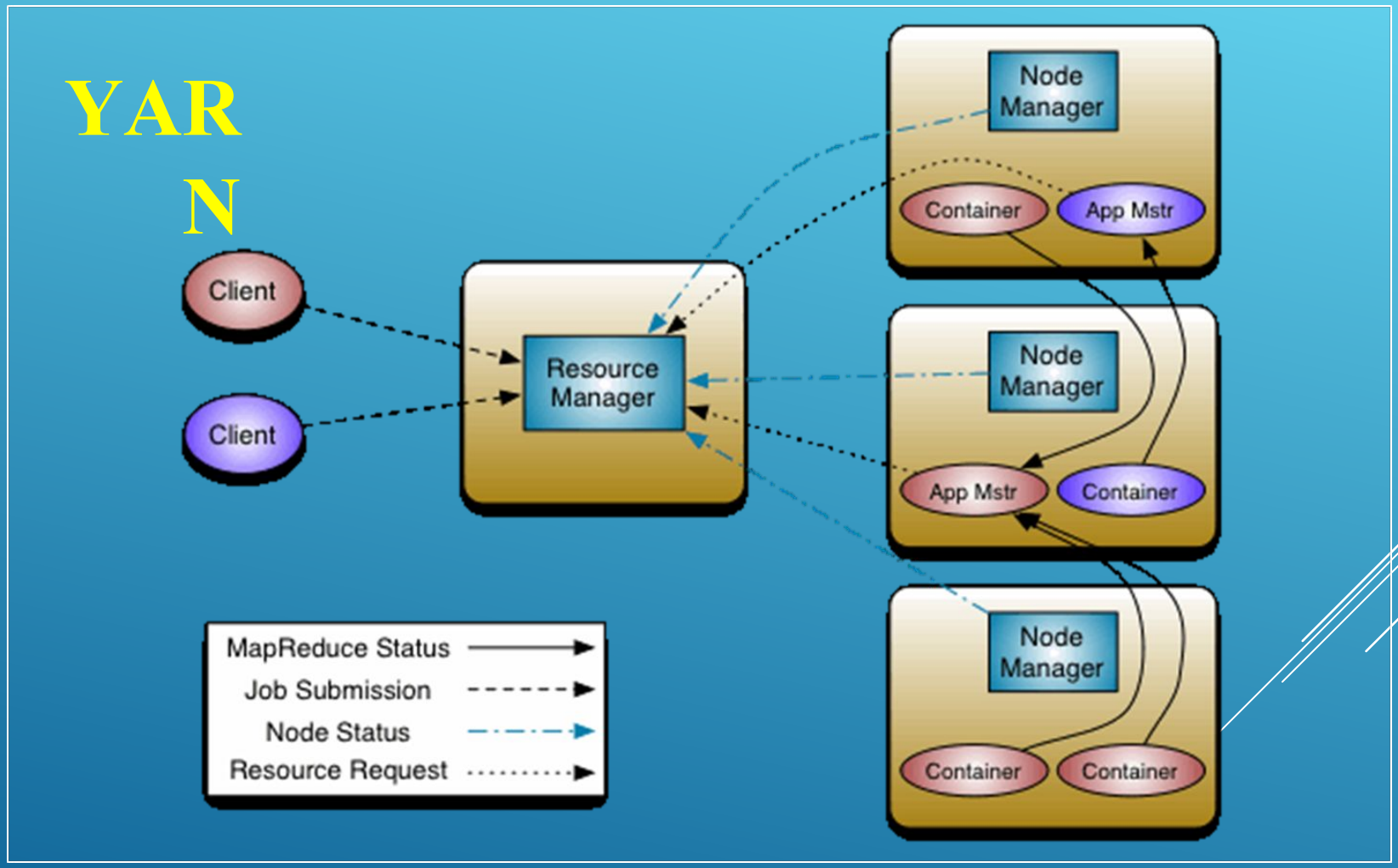


MR VS. YARN ARCHITECTURE



- **YARN** : **Y**et **A**nother **R**esource **N**egotiator
- **MR** : MapReduce

YARN


Client
Client
Resource Manager
Node Manager
Container
App Mstr
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request